DOCUMENT RESUME

ED 403 275 TM 025 949

AUTHOR Powers, Donald E.; Potenza, Maria T.
TITLE Comparability of Testing Using Laptop and Desktop Computers.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-15
PUB DATE Apr 96
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Students; Comparative Analysis; *Computer Assisted Testing; Higher Education; Mathematics Tests; *Microcomputers; *Test Construction; *Testing Problems; Test Results; Verbal Tests; Writing Tests
IDENTIFIERS *Graduate Record Examinations; *Laptop Computers

ABSTRACT

The degree to which laptop and standard-size desktop computers are likely to produce comparable test results for the Graduate Record Examination (GRE) General Test was studied. Verbal, quantitative, and writing sections of a retired version of the GRE were used, since it was expected that performance on reading passages or mathematics items might be affected by monitor size and the additional scrolling needed for a laptop. Subjects were 201 paid volunteer graduate students and upper-level undergraduates on 9 university campuses, all of whom had at least minimal typing skills. Usable data were available for 200 subjects on the verbal and quantitative tests and 199 on the writing portion. All subjects participated with both types of computer. Analyses of test scores indicate that only performance on the essay section was affected by the type of computer used, and only for the first of two essays. The reason for the interaction of mode with order of testing was not evident, although fatigue may have modified the results of the second essay, which was written after earlier testing. The laptop model used in this study appears likely to yield results that are comparable to those obtained with a standard desktop model. (Contains seven tables and two references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-96-15

ED 403 275

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

COMPARABILITY OF TESTING USING LAPTOP AND DESKTOP COMPUTERS

Donald E. Powers
Maria T. Potenza

Educational Testing Service
Princeton, New Jersey
April 1996

TM025949

2

BEST COPY AVAILABLE

Comparability of Testing Using Laptop and Desktop Computers

Donald E. Powers

Maria T. Potenza

Comparability of Testing Using Laptop and Desktop Computers

Standardized testing is predicated on the belief that, by presenting each examinee with exactly the same (or strictly equivalent) tasks given under identical conditions, test results are comparable across examinees and thus fair to every test taker. When important aspects of standardization are breached -- for example when some test takers are permitted more time than others or when some examinees receive a more difficult version of a test -- the resulting test scores may not be comparable and therefore may be potentially unfair to some examinees.

Generally, for paper-and-pencil tests, minor differences from one test administration to another in the characteristics of the testing medium are assumed to matter very little, if at all. When tests are computer-based, however, standardization is more complex: there are many (possibly important) ways in which computer administrations may differ, thus requiring greater vigilance by test makers in order to ensure that comparability is maintained. Examinees may, for instance, be differentially familiar with various models of computers and with different methods of inputting responses (keyboard vs. mouse, for example). Moreover, the use of different computers may be associated with variability in such factors as the clarity with which text appears to examinees and the rate at which test questions are presented.

As major computer-based testing programs such as the Graduate Record Examinations (GRE) move increasingly into international markets, it is likely that alternative delivery methods will need to be considered. Delivery systems that are highly portable, such as laptop computers, have much appeal. To the extent that delivery may differ from site to site, some assurance will be needed that any necessary or unavoidable differences will not impinge on test performance. This assurance is in fact dictated by professional standards for computer-based testing:

> Any departure from the standard equipment, conditions, or procedures, as described in the test manual or administrative instructions [for computer-based administration], should be demonstrated not to affect test scores appreciably. Otherwise, appropriate calibration should be undertaken and documented (Committee on Professional Standards and Committee on Psychological Tests and Assessment, 1986, p. 10).

This study was designed to evaluate the degree to which laptop and standard-size desktop computers are likely to produce comparable test results for the GRE General Test. In the event that scores exhibit some degree of noncomparability, the second objective was to suggest ways in which noncomparability might be reduced or eliminated.

Method

Instruments

A retired version of the GRE General Test was identified for use in the study. Two abbreviated "forms" of the test were constituted by combining different sections from the six separately-timed sections comprising the total test. Only verbal and quantitative sections, not analytical ones, were used for this study, as the verbal and quantitative sections were believed to contain a sufficient variety of item types to allow an adequate test of the research hypothesis. In particular, these sections contained reading comprehension items, which are associated with the longest passages of text in the examination, and quantitative reasoning items, which employ mathematical notation and graphical stimuli. Performance on the former item type might be affected by the size of the monitors and the additional scrolling needed for laptop

administration, it was conjectured. For the latter item type, it seemed possible that the appearance might depend on the particular hardware used, especially the size of the monitor. In addition, two essay topics that had been pretested for the new GRE writing test were identified and included in the study in order to provide a more rigorous test (than was possible with multiple-choice questions) of the extent to which differences in keyboards might affect test performance.

Each of the two test forms contained one verbal followed by one quantitative section of the full test. To each of these forms was added one of the two essay topics. Finally, a questionnaire was administered to obtain a variety of background information about examinees and to get their perceptions of various aspects of testing with both the laptop and desktop computers.

Equipment

The desktop computers used in the study were Compaq 486 models (25mhz) with color monitors, 8 MB RAM and 527MB harddrives. The laptop was a 33mhz Toshiba model (T6400 DXC) with a smaller harddrive (200 MB). Its dimensions in inches were 10.5 (length) x 15.4 (width) x 4.1 (height). The screen resolution was 640 x 480 and the display matrix was active. The size of the monitors was 14 inches for the desktop models and 10.4 for the laptops. The desktop keyboards were standard IBM configurations; the laptop model included an integrated numeric keyboard. When the study was initiated, these models were the ones thought most likely to be used for actual test administrations.

Subjects

Subjects were 201 paid volunteers recruited by graduate research assistants on nine university campuses (Johns Hopkins University, Stanford University, University of California at Berkeley, University of Colorado, University of Illinois, University of Massachusetts, University of Ottawa, University of Texas -- Pan American, and Utah State University). The research assistants were directed to target first-year graduate students and upper-division undergraduates, to obtain a mix in terms of major area of study, to secure a roughly equal balance with respect to gender, and to obtain at least some representation of ethnic minority students in the sample. Having at least minimal typing skills was the only firm prerequisite to participating in the study, so that essays could be wordprocessed. The eventual number of subjects from each school ranged from 14 to 30.

Design

Subjects were assigned sequentially to one of four study conditions determined by the order of administration of the desktop and laptop versions and the two different test forms. Thus, the order of administration of both test form and mode of testing was completely counterbalanced. After taking as much time as needed to complete a tutorial explaining testing procedures, all subjects took both of the abbreviated forms of the test, one on a laptop computer and the other on the desktop model. These individual testing sessions were conducted immediately after one another, with a short break between them. Before commencing to work on the second test, each subject was given time to "warm up" on the second keyboard in order

to become accustomed to the size and possibly different feel of the smaller (or larger) keyboard. Subjects completed the questionnaire immediately after testing.

Analyses

To take advantage of the fact that each subject in effect served as his/her own comparison, a repeated measures analysis of variance was used to analyze test scores. Three separate analyses were run, in which the dependent variable was the score on the GRE test section -- either verbal or quantitative -- or the score on the essay. The repeated measures were test form and mode of testing.

Although each section of the GRE General Test is developed to the same specifications as its like-content counterpart and is therefore assumed to be parallel, no formal equating of test sections is conducted. Therefore, in order to pool test results across two test "forms" for this study we first scaled each form to have a mean of 0 and a standard deviation of 1.

Results

The Sample

Useable test data were available for 200 subjects for the verbal and quantitative tests and for 199 subjects for the writing portion. Table 1 describes the study sample in terms of several relevant background characteristics. When compared with the GRE test-taking population (e.g., Wah & Robinson, 1990), the sample appears to be reasonably representative of people who take the GRE General Test. It is noteworthy, perhaps, that a majority of the sample (61%) reported wearing corrective lenses, a condition that, as discussed later, may have some relevance for the comparison of test performances on laptop and desktop computers. Information on another relevant characteristic of the sample -- frequency of use of desktop and laptop computers -- is given in Table 2. As can be seen, as a group, the sample was more experienced with desktop than with laptop computers: the vast majority of subjects (94%) reported regular or routine use of desktop computers, but far fewer (21%) reported using laptop computers this frequently. In fact, more than a quarter of the sample said that, before participating in the study, they had never used a laptop computer. Participants reported having varying levels of keyboarding skills. Most described their skills as being relatively good -- about average (35%), a little above average (40%), or far above average (12%). Far fewer characterized themselves as being a little below average (10%) or far below average (2%). Relatively few (18%) of the study participants said that they had taken the GRE General Test previously, and 89% of these reported their scores to us. For this subsample the mean GRE General Test scores were 549 (sd=122), 646 (sd=140), and 633 (sd=144), respectively, for the verbal, quantitative, and analytical portions of the test. By contrast, the means for 1991-92 GRE General Test takers were 485 (sd=118), 553 (sd=139), and 536 (sd=129), respectively.

Examinee Perceptions

When asked whether they found test taking to be easier on one kind of computer than the other, more than a third of the sample (36%) felt that their experience was "about the same on both" computers. Nearly half (48%), however, said that test taking was easier on the desktop model. A minority (15%) felt that it was easier to take the test on the laptop model.

In order to pinpoint the specific features that differentiated test taking on the two kinds of computers, we asked subjects to rate the adequacy of the desktop and laptop computers with respect to a number of possibly salient characteristics. A summary of responses to this query is given in Table 3. As expected, there were no differences between computers with respect to the perceived adequacy of either the speed of presentation of questions or the useability of the mouse, both of which were regarded as adequate or more than adequate by approximately 95% of the sample. Each of the other features listed in Table 3 was judged as being somewhat less adequate on the laptop computer than on the desktop model. The largest discrepancies pertained to the size of the keyboard, the size of the screen, and the touch/feel of the keyboard. Even for these features, however, a majority of study participants believed them to be at least adequate (59%, 68%, and 64%, respectively, and relatively few rated them as being deficient. (Only 6%, 5%, and 9% of respondents, respectively, judged these features to be inadequate.) About one fourth to one third of the sample did, however, believe that these features were marginal on the laptop model.

To obtain additional insight into examinees' perceptions of the differences between the two kinds of computers, we asked participants to describe problems that they experienced with test taking on one computer but not on the other. The comments made most frequently (by 11% of the sample) pertained to the unfamiliar location/position of the keys on the laptop keyboard. Specifically, the proximity of the "home" and the "backspace" keys was problematical. Typical of the responses were:

> The keyboard layout of the laptop is not normal. A lot of times I [tried to] backspace but got the home key instead.
>
> It was much harder to do corrections for the essay on the laptop because of the close proximity of the arrow keys. I wasted a lot of time because I kept hitting the key I thought was the arrow.
>
> I'm more used to a desktop and tend to press the wrong key on the laptop keyboard due to a different arrangement.

The second most frequent kind of comment, offered by nearly 10% of the sample, concerned the closeness of the keys on the laptop model:

> [I was] not used to the closeness of the keys.
>
> Keys were so close together I sometimes hit the wrong one while typing.
>
> While writing the essay, the keys are very small so many more spelling errors were made.

About 7% of the sample suggested that the keyboard was too small generally, about 6% made general comments about the laptop keyboard "not being as good," and about 5% commented that the laptop keyboard was too sensitive to the touch. (As proof that it is usually impossible to please everyone, we note that several participants told us that the desktop keyboards were too noisy, too stiff, or too large and "bulky.")

The next most frequent comments involved the laptop screen. About 8% of study subjects suggested that it was too small, and about 9% commented on problems with glare or with a lack of clarity. Similar comments were very rare for the desktop model, although about 4% of respondents were somewhat unhappy with the position or angle of the desktop monitor.

Because the various kinds of questions on the GRE General Test differ with respect to both appearance and type of response required, we asked study participants to rate the difficulty of each item type when delivered by laptop and desktop computers. Table 4 shows that, with

two exceptions, none of the GRE item types were perceived as being more difficult on one computer than the other: approximately 90% of the sample judged each item type to be equally difficult on both kinds of computers. The exceptions were reading comprehension and, especially, essay writing, both of which were viewed more often as being more difficult on the laptop than more difficult on the desktop. Even for these two item types, however, a majority (67%) of respondents believed that reading comprehension items were equally difficult on both types of computers, and a plurality (46%) felt that essay writing was equally difficult on both.

When asked to tell us why some question types seemed more difficult than others, study subjects most often mentioned keyboard-related problems for the essay, and screen-related problems for reading comprehension questions. The inability to view an entire reading passage on a single screen was a concern for about 8% of the study participants, either for reading comprehension items or for problems involving charts, graphs, or figures. A related issue, mentioned by about 2% of the sample, was the need for more scrolling on the laptop than on the desktop computer.

Finally, subjects were asked for their suggestions on how to make test taking easier on either laptop or desktop computers. The most frequently offered suggestion (by about 8% of the sample) was to improve the laptop keyboard – by enlarging it and by rearranging the keys. The next most frequent advice (from about 7% of the sample) was to reduce or eliminate the need to scroll for reading comprehension questions as well as for other kinds of questions. This advice pertains to both types of computers. Finally, about 5% of the study respondents suggested that we allow examinees to omit questions and review them later. A variety of other recommendations were made far less frequently.

Test Performance

Table 5 shows, in standard deviation (z score) units, the mean test scores by mode and order of administration for each test. (As mentioned earlier, raw scores for each test have been converted to z-score units in order to pool test results from each of the two test "forms" that was administered.) The correlations between scores based on desktop administration and those based on laptop administration were .87, .82, and .69 for the verbal, quantitative, and writing portions of the test, respectively. These strong correlations suggest that examinees were rank ordered in a very similar manner regardless of the mode of test delivery. For the verbal and quantitative measures these correlations are about as high as could be expected according to the likely reliability of the individual sections. The lower correlation for writing is about what is typically observed for correlations between two different essay topics. (The interreader correlations for the two essays used in the study were .84 and .86.)

Table 6 displays the results of the repeated measures analyses of variance, and Table 7 gives the sizes of the effects of mode and order of test administration for each test. As can be seen from Table 6, there were no significant effects of either mode or order for either the verbal or the quantitative measures (although for the quantitative measure the effect of order was very nearly significant at the .05 level). For writing, however, both the order and the mode-by-order terms were statistically significant, suggesting that there was an effect of mode of test delivery, which differed according to whether the test was given first or second.

Table 7 shows the small, nonsignificant effects of mode for the verbal and quantitative tests, and the somewhat larger (though still relatively small) effects of order. The latter, revealing slightly lower scores when tests were taken second, is suggestive of a fatigue effect. This interpretation seems quite plausible, as when they encountered the second verbal test, participants had already devoted one hour and 40 minutes to test taking. Fatigue may have come into play to an even greater degree for the second quantitative test and even more still for the second essay, which was the final task to be completed for the study.

The results for the essay writing portion of the test are somewhat more complex than for the verbal and quantitative measures, as there was a significant ($p<.05$) interaction between mode and order of testing. This interaction may have been the result of a carryover effect. Study participants who wrote first on desktop computers (and then laptops) performed less well on the second essay than the first. However, participants who wrote on the laptops first did not exhibit a lower performance on the second (desktop) essay. It seems possible that composing first on the larger, standard desktop keyboard may have made it more difficult to adjust to the smaller laptop keyboard. The effect size of .24 for the first essay, though small, was statistically significant.

Discussion

That all subjects participated in both of the major study conditions, i.e., testing both on laptop and on desktop computers, was both a strength and a limitation of the study. Comparisons between treatments could be made with a high degree of precision under this design. However, both examinee perceptions, as well as their test performances, may have been influenced by their involvement in both experimental conditions. For instance, it seems likely that subjects might have reported less dissatisfaction with the size of the laptop keyboard if they had used only this keyboard during the study. Thus, by explicitly incorporating a comparative aspect into the study design we invited a relatively stringent evaluation of the laptop. In addition, we suspected that the study design might exaggerate any differences between test performances on the two types of computers, insofar, for example, that having taken a test on a full-size keyboard might make immediate subsequent testing on a smaller keyboard even more difficult than it normally would be. The comparisons made here were undoubtedly affected also by the nature of the sample, whose members were, generally speaking, relatively inexperienced with laptop computers.

Nonetheless, analyses of test scores revealed that of the three GRE measures that were studied, only performance on the essay portion of the test was affected by the mode of testing, but only for the first of two essays written by participants. The reason for this interaction (i.e., mode by order of testing) was not readily apparent. One conjecture is that fatigue may have modified the results for the second essay that was written, as second essays were drafted after 2 hours and 40 minutes of earlier testing. It is not clear, however, how fatigue would account for the particular pattern of results that was observed here.

On the basis of the results reported here, it would appear that the particular model of laptop computer used in this study is likely to provide test results that are comparable to those obtained with a standard-size desktop model -- at least for the standard multiple-choice questions used in the GRE General Test, but probably not for the portion of the test that

involves essay writing. With the substitution of a standard-size keyboard, however, it seems very likely that comparable results could be attained for this section of the test also.

Finally, the study results suggest that, as future delivery systems are developed, attention could profitably be given to features that enhance the presentation of certain test item types, such as reading comprehension questions. Even though performance on these item types did not contribute to differences between modes of testing, these items were perceived as more difficult when presented on laptop than on desktop computers. This perception seems worthy of attention, even if it does not translate to actual score differences between modes of test delivery. In addition, further monitoring of comparability (especially for international students) will be desirable, and plans are now being established to carry this out.

References

Committee on Professional Standards and Committee on Psychological Tests and Assessment. (1986). Guidelines for computer-based tests and interpretations. Washington, DC: American Psychological Association.

Wah, D.M. & Robinson, D.S. (1990). Examinee and score trends for the GRE General Test: 1977-78, 1982-83, 1986-87, 1987-88. Princeton, NJ: Educational Testing Service.

Table 1

Description of Study Sample

Sex (% female)	57
English best language (%)	93
Age (%):	
under 21	16
21-25	73
26-30	5
over 30	6
Undergraduate Major Field (%):	
Natural Science	21
Engineering	17
Social Sciences	28
Humanities and Arts	14
Education	8
Business	3
Other	7
Unknown	3
Wear corrective lenses (%)	61
Physical condition making computerized testing difficult (%)	1
Ethnicity (%):	
Black or African American	8
Mexican American or Chicano	10
Asian American or Pacific Islander	19
Puerto Rican	1
Other Hispanic	6
White	50
Other	4

Table 2

Frequency of Use of Desktop and Laptop Computers by Study Participants

Frequency of Use	Desktop	Laptop
Never before today's test	<1%	28%
Rarely (only a few times in the last five years)	5	52
Regularly (sometime each week)	35	14
Routinely (almost daily)	59	7

Note. One participant did not respond regarding experience with desktops and 24 did not respond regarding experience with laptops.

Table 3

Perceived Adequacy of Selected Feature of Laptop and Desktop Computers for Testing

Feature	Computer	Adequacy			
		More than adequate	Adequate	Marginal	Inadequate
Size of keyboard	Desktop	59%	38%	2%	<1%
	Laptop	21	38	**35**	**6**
Touch/feel of keyboard	Desktop	43	44	10	3
	Laptop	32	32	**27**	**9**
Size of screen	Desktop	53	43	3	<1
	Laptop	23	45	**26**	**5**
Size of print	Desktop	48	51	1	0
	Laptop	23	63	13	1
Clarity of print	Desktop	49	48	2	<1
	Laptop	27	53	**18**	2
Clarity of figures/ diagrams	Desktop	47	49	3	1
	Laptop	32	49	**16**	2
Speed of presentation	Desktop	46	49	5	<1
	Laptop	46	49	3	1
Mouse useability	Desktop	53	44	3	<1
	Laptop	55	39	5	<1

Table 4

Perceptions of Difficulty of Various Kinds of Questions

when Presented on Desktop vs. Laptop

	Perceived Difficulty		
Question Type	More difficult on laptop	Equally difficult on both	More difficult on desktop
Analogies	6%	91%	2%
Antonyms	2	94	3
Sentence completions	9	88	4
Reading comprehension	28	67	5
Arithmetic/algebra	7	90	4
Geometry	9	86	6
Quantitative comparisons	7	91	3
Problem solving	7	91	3
Essay writing	43	46	11

Table 5

GRE Test Performance by Order of Test Administration and Mode of Delivery

Mode	Order	
	First	Second
	Verbal	
Desktop	.13	-.08
Laptop	.08	-.14
	Quantitative	
Desktop	.17	-.13
Laptop	.09	-.14
	Writing	
Desktop	.18	-.06
Laptop	-.06	-.04

Note. Table entries are z - score means. Standard deviations ranged from .9 to 1.1.

Table 6

Analyses of Variance Results by Test Section

Factor	Source	df	MS	F	sig of F
		Verbal			
Between Subjects	Order	1	4.39	2.37	n.s.
	Within Cells	198	1.85		
Within Subjects	Mode	1	.32	2.45	n.s.
	Order by Mode	1	.00	.02	n.s.
	Within Cells	198			
		Quantitative			
Between Subjects	Order	1	6.89	3.85	.051
	Within Cells	198	1.79		
Within Subjects	Mode	1	.20	1.11	n.s.
	Order by Mode	1	.12	.65	n.s.
	Within Cells	198	.18		
		Writing			
Between Subjects	Order	1	1.15	.68	n.s.
	Within Cells	196	1.69		
Within Subjects	Mode	1	1.18	3.89	.05
	Order by Mode	1	1.59	5.21	<.05
	Within Cells	196	.30		

Table 7

Effect Estimates for Mode and Order of Test Administration

	Test Section		
Effect	Verbal	Quantitative	Writing
Mode	.06	.04	.11*
Order	.21	.26	.11*

Note. Effect estimates are given in standard deviation units.

*There was a significant interaction of mode and order of testing for the writing test. The effect estimate when tests were given first was .24; when given second, the effect was -.02. Positive numbers indicate that scores were higher on the desktop than laptop computers and higher when tests were given first than second.

Acknowledgments

The following people all contributed to the study reported here:

Stef Bogdan
Lyle Brenner
Karen Draney
Rob Durso
Mary Fowles (and the essay readers)
Lenora Green
Candus Hedberg
Melissa Kay
Charlie Lewis
Craig Mills
Philip Moberg
Bill Nemceff
Phil Oltman
Robin Pergament
Vittorio Puente
Pam Rice
David Schmidt
Deborah Schnipke
Sharon Slater
Barbara Soriano
Len Swanson
Fang Tian
Laurie Van Sant
Timothy Weston
Lou Woodruff
Tom Yi
Ruth Yoder

Thanks to all of you.